Towards efficient search tools for biomedical databases: Characterizing user search habits and recognizing their information needs

Rezarta Islamaj Doğan, G. Craig Murray, Aurélie Névéol and Zhiyong Lu

National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD 20894.

ABSTRACT

Log analysis is a useful way to understand user needs and their search habits for improving information retrieval systems. This study provides insight into PubMed® users search habits. We analyzed more than 100 million PubMed user queries, abstract views and full text views. Additionally, 10,000 user queries were manually reviewed and categorized into several search requests categories. Our analysis revealed unique characteristics for biomedical information searches. PubMed users are persistent in seeking information; they reformulate their queries 47% of the time and browse results 44% of the time. On average, PubMed users click on 4 citations per query. And, after an abstract view, the full text of the article is requested 29% of the time. PubMed users' decisions to click on articles are influenced by the result set size.

Biomedical information queries are short, thus each word has significant impact for results. The three most popular types of search are: search by author name (36%), disease name (20%) and gene/protein name (19%). Queries that contain bibliographic words, such as author names, have a smaller result set than average. For the other cases, we collected the articles clicked on average at least once per user per day and matched them with the queries that led to those clicks. We studied the popular query words to help identify the users' information needs. These words have usually a high TF-IDF weight compared to the other words in the abstract and are mostly found in the title.

Such an analysis provides useful insight for improving retrieval quality of PubMed and other biomedical information systems. It suggests that specialized techniques might be more desirable than traditional ones, and it may also impact research on the fields of document indexing, categorization and summarization.

ACKNOWLEDGEMENTS

This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.

REFERENCES

Islamaj Dogan R, *et al*. (2009) Understanding PubMed® user search behavior through log analysis. Database. Vol. 2009: bap018;